

Estimation of scale functions to model heteroscedasticity by support vector machines

Robert Hable

University of Bayreuth
Department of Mathematics
D-95440 Bayreuth
robert.hable@uni-bayreuth.de

Andreas Christmann

University of Bayreuth
Department of Mathematics
D-95440 Bayreuth
andreas.christmann@uni-bayreuth.de

November 9, 2011

Abstract

A main goal of regression is to derive statistical conclusions on the conditional distribution of the output variable Y given the input values x . Two of the most important characteristics of a single distribution are location and scale. Support vector machines (SVMs) are well established to estimate location functions like the conditional median or the conditional mean. We investigate the estimation of scale functions by SVMs when the conditional median is unknown, too. Estimation of scale functions is important e.g. to estimate the volatility in finance. We consider the median absolute deviation (MAD) and the interquantile range (IQR) as measures of scale. Our main result shows the consistency of MAD-type SVMs.

1 Introduction

Let P be the distribution of a pair of random variables (X, Y) with values in a set $\mathcal{X} \times \mathcal{Y}$ where X is an input variable and Y is a real-valued output variable. The goal in regression problems is to derive statistical conclusions on the conditional distribution of Y given $X = x$. Generally, location and scale are considered as the two most important characteristics of a distribution and estimating these quantities is one of the main topics in statistics.

Regularized empirical risk minimization [26, 27, 18, 8] using the kernel trick proposed by [19] and the special case of *support vector machines* (SVMs) [26, 6, 18, 23] are well established methods in order to estimate the location of the conditional distribution of Y given $X = x$. For an i.i.d. sample $D = ((X_1, Y_1), \dots, (X_n, Y_n))$ drawn from P , the SVM-estimator is defined by

$$f_{L,D,\lambda} = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^n L(Y_i, f(X_i)) + \lambda \|f\|_H^2, \quad (1)$$

where L is a loss function, H is a certain space – a so-called *reproducing kernel Hilbert space* (RKHS) – of functions $f : \mathcal{X} \rightarrow \mathbb{R}$, and $\lambda \in (0, \infty)$ is a regularization parameter in order to prevent overfitting. We refer to [27, 18, 2, 8, 23] for the concept of an RKHS. There are a number of different quantities which describe the location of a single distribution and which can be estimated by SVMs by choosing

a suitable loss function. The conditional mean function $g(x) := \mathbb{E}_P[Y|X = x]$, $x \in \mathcal{X}$ can be estimated by using the least-squares loss $L_{LS}(y, t) = (y - t)^2$ and the τ -quantile function $g(x) := f_{\tau, P}^*(x)$, $x \in \mathcal{X}$, (see (2) below) by using the τ -pinball loss function

$$L_{\tau\text{-pin}}(y, t) = \begin{cases} (1 - \tau) \cdot (t - y) & \text{if } y - t < 0, \\ \tau \cdot (y - t) & \text{if } y - t \geq 0, \end{cases} \quad (y, t) \in \mathcal{Y} \times \mathbb{R},$$

see [15, 14, 20, 25]. The choice $\tau = 0.5$ leads to an estimate of the *median function*

$$f_{0.5, P}^*(x) := \text{median}_P(Y|X = x), \quad x \in \mathcal{X}.$$

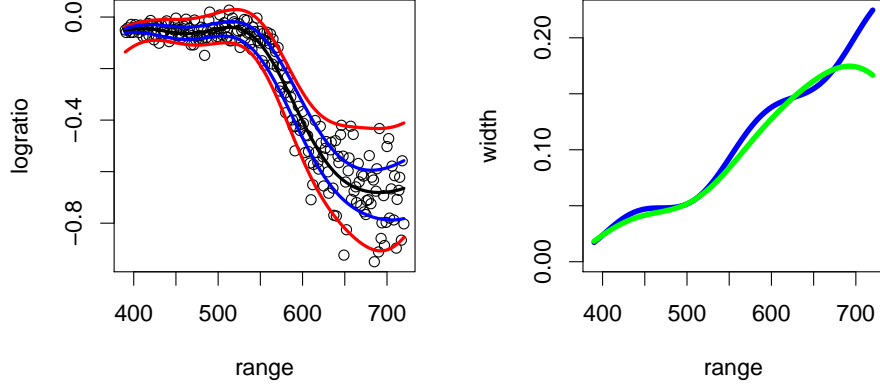
The goal of this paper is to investigate two methods to estimate the variability of the conditional distributions of Y given $X = x$ for $x \in \mathcal{X}$ via *scale functions*. Estimation of heteroscedasticity is interesting in many areas of applied statistics, e.g., for the estimation of volatility in finance. To fix ideas, let us *illustrate* what we mean by scale function estimation by considering a small data concerning the so-called LIDAR technique. LIDAR is the abbreviation of Light Detection And Ranging. This technique uses the reflection of laser-emitted light to detect chemical compounds in the atmosphere. We consider the logarithm of the ratio of light received from two laser sources as the output variable $Y = \text{logratio}$, whereas the single input variable $X = \text{range}$ is the distance traveled before the light is reflected back to its source. We refer to [17] for more details on this data set. A scatterplot of the data set consisting of $n = 221$ observations is shown in the left subplot of Figure 1 together with the fitted quantile curves based on SVMs using the pinball loss function for $\tau \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$ and the Gaussian RBF kernel $k(x, x') := \exp(-\gamma\|x - x'\|_2^2)$ for $x, x' \in \mathcal{X}$. By looking at the estimated median function (i.e., the black curve in the middle of the left subplot), we clearly see a nonlinear relationship between both variables which is almost constant for values of **range** below 550 and decreasing for higher values of **range**. However, there is also a considerable change of the variability of **logratio** given **range**: the variability is relatively small for small values of **range**, but much larger for large values of **range**. This becomes obvious by looking at the other estimated quantile curves in the left subplot or by looking at the right subplot of Figure 1 which shows the estimated width of intervals covering at least 50% of the mass of $P(Y|x)$. In this simple example we can just look onto the 2-dimensional plot to realize this kind of heteroscedasticity of the conditional distribution of Y given $X = x$. However, this is obviously no longer possible if the input space \mathcal{X} is a high-dimensional Euclidean space or an abstract metric space. Hence an automatic and non-parametric method to model and to estimate such kind of variability becomes important. Therefore, this article investigates how two classical scale quantities of the conditional distribution of Y given $X = x$ can be estimated by use of SVMs. Such scale functions $g : \mathcal{X} \rightarrow [0, \infty)$ are quite common in a heteroscedastic model like $P(Y|x) = f(x) + g(x)\varepsilon$, where f denotes the location function and ε denotes the stochastic error term. Note, that we will *not* assume such a specific model. As in case of location, there are several well established quantities which describe the scale, e.g., [10, Chap. 5]

- (i) the *variance function*: $g(x) = \text{Var}_P(Y|X = x)$, $x \in \mathcal{X}$,
- (ii) the *median absolute deviation from the median (MAD) function*:
 $g(x) := \text{MAD}_P(Y|X = x) := \text{median}(|Y - f_{0.5, P}^*(x)| | X = x)$, $x \in \mathcal{X}$,
- (iii) the *interquantile range (IQR) function for quantiles $\tau_1 < \tau_2$* :
 $g(x) := \text{IQR}_{\tau_1, \tau_2}(Y|X = x) := f_{\tau_2, P}^*(x) - f_{\tau_1, P}^*(x)$, $x \in \mathcal{X}$.

Note that these three quantities are *not* directly comparable. However, $\text{IQR}_{0.25, 0.75}$ and 2 times MAD are both quantities for the width of an interval covering at least 50% of the probability mass of $P(Y|x)$. There is a vast literature on the estimation of scale functions, often based on special parametric dispersion models, see, e.g., [11, 21, 12], and for a wavelet thresholding approach for univariate regression models we refer to [3].

In this article, we consider the MAD function and the IQR function and show how both can be consistently estimated in a purely nonparametric manner with SVMs. In case of the MAD, we estimate the unknown median function $f_{0.5, P}^*$ by an SVM $f_{L_{0.5\text{-pin}}, D, \lambda}$ and calculate the estimated absolute residuals $R_i := |Y_i - f_{L_{0.5\text{-pin}}, D, \lambda}(X_i)|$ in a first step. In a second step, we estimate the

Figure 1: Illustration of the estimation for scale functions by SVMS for the LIDAR data set. Left subplot: data set, estimated quantile functions with SVMS for $\tau = 0.5$ (black), $\tau = 0.25$ and $\tau = 0.75$ (both in blue), $\tau = 0.05$ and $\tau = 0.95$ (both in red). Right subplot: Estimated width of the intervals covering 50% of the mass of $P(Y|x)$. IQR-type SVM (blue) using $(\tau_1, \tau_2) = (0.25, 0.75)$ and 2 times the MAD-type SVM (green).



conditional median of the absolute residuals by the SVM based on a smoothed version of the $\frac{1}{2}$ -pinball loss defined in (4) below for the pairs of random variables (X_i, R_i) . The resulting estimator is called MAD-type SVM and it is shown in Subsection 2.1 that it is risk-consistent (up to any predefined $\varepsilon > 0$) even though (i) the estimation in the second step cannot be based on the true residuals but has to be based on the estimated residuals because the true median function is unknown and (ii) the random variables (X_i, R_i) are not i.i.d. In case of the IQR $f_{\tau_2, P}^* - f_{\tau_1, P}^*$, we respectively estimate the τ_j -quantile function $f_{\tau_j, P}^*$ by use of the τ_j -pinball loss so that we get $f_{L_{\tau_2\text{-pin}, D}, \lambda_2} - f_{L_{\tau_1\text{-pin}, D}, \lambda_1}$ as an estimate, which we call IQR-type SVM. As this is the difference of two standard SVMs, we can carry over many well-known facts on SVMs in this case in Subsection 2.2. In both cases, available software, e.g., the R-package `kernlab` [13] or the C++ implementation `mySVM` [16], can be used since we essentially have to calculate SVMs for pinball losses.

The rest of the paper has the following structure. Section 2 contains with Theorem 2.2 our main result. Section 3 contains not only the proof of this theorem, but also gives two new consistency results in the L_1 -sense for SVMs based on the pinball loss, see Lemma 3.1 and Theorem 3.2. Although we need these results in our proof of Theorem 2.2, we think that they are interesting in its own, because they improve earlier consistency results of SVMs which showed the weaker kind of convergence in probability, see [23, Cor. 3.62, Thm. 9.7(i)].

2 Main results

The following assumptions and notations are used throughout the whole article.

Assumption 2.1 Let \mathcal{X} be a complete separable metric space, e.g. $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{Y} \subset \mathbb{R}$ be closed. For $j \in \{1, 2\}$, let $k_j : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a bounded continuous kernel with $\|k_j\|_\infty := \sup_{x \in \mathcal{X}} (k_j(x, x))^{1/2} < \infty$. Its corresponding reproducing kernel Hilbert space (RKHS) is denoted by H_j , its corresponding canonical feature map is denoted by Φ_j , and it is assumed that each H_j is dense in $L_1(\mu)$ for every $\mu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$.

$\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ denotes the set of all Borel probability measures on $\mathcal{X} \times \mathcal{Y}$. The unknown joint distribution of (X, Y) is denoted by $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and $D = ((X_1, Y_1), \dots, (X_n, Y_n))$ is an i.i.d. sample drawn from P . Let $P_{\mathcal{X}}$ denote the distribution of X , let $\mathcal{L}_0(\mathcal{X})$ denote the set of all Borel measurable functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and let $L_1(P_{\mathcal{X}})$ denote the set of all $P_{\mathcal{X}}$ -integrable functions $f : \mathcal{X} \rightarrow \mathbb{R}$. We

define the τ -quantile function as the (perhaps set-valued) function

$$\mathcal{X} \rightarrow 2^{\mathbb{R}}, \quad x \mapsto F_{\tau, P}^*(x) := \{t \in \mathbb{R} : P((-\infty, t] | x) \geq \tau \text{ and } P([t, \infty) | x) \geq 1 - \tau\}. \quad (2)$$

We make the standard assumption that $F_{\tau, P}^*(x)$ are singletons and hence we write $f_{\tau, P}^*(x) : \mathcal{X} \rightarrow \mathbb{R}$ instead, see [22, 24].

2.1 MAD-type estimation

We would like to estimate the MAD function given by $g(x) = \text{MAD}_P(Y|X = x) = \text{median}_P(|Y - f_{0.5, P}^*(x)| | X = x)$ where $f_{0.5, P}^*$ is the median function. First, we estimate the median function $f_{0.5, P}^*$. For a random sample $D = ((X_1, Y_1), \dots, (X_n, Y_n))$ drawn from P , the SVM-estimator for $f_{0.5, P}^*$ is

$$f_{L_{0.5-\text{pin}}, D, \lambda_{1, n}} = \arg \inf_{f \in H_1} \left(\frac{1}{n} \sum_{i=1}^n L_{0.5-\text{pin}}(Y_i, f(X_i)) + \lambda_{1, n} \|f\|_{H_1}^2 \right),$$

$\lambda_{1, n} \in (0, \infty)$, and H_1 is an RKHS. Then, we can estimate the conditional median of the absolute residuals $|Y - f_{0.5, P}^*(x)|$ by use of the *estimated* absolute residuals. Let us define

$$\tilde{g}_{D, n} = \arg \inf_{g \in H_2} \left(\frac{1}{n} \sum_{i=1}^n L_{\varepsilon}(|Y_i - f_{L_{0.5-\text{pin}}, D, \lambda_{1, n}}(X_i)|, g(X_i)) + \lambda_{2, n} \|g\|_{H_2}^2 \right), \quad (3)$$

where, for some small predefined number $\varepsilon > 0$, the loss function L_{ε} defined by

$$L_{\varepsilon}(y, t) = \frac{1}{2}(y - t) - \varepsilon \log(2\Lambda(\frac{y-t}{\varepsilon})) = L_{0.5-\text{pin}}(y, t) - \varepsilon \log(2\Lambda(\frac{|y-t|}{\varepsilon})), \quad (4)$$

is an ε -smoothed version of the pinball loss function for $\tau = 0.5$, $\Lambda(r) = 1/(1 + e^{-r})$ for every $r \in \mathbb{R}$, $\lambda_{2, n} \in (0, \infty)$, and H_2 is an RKHS. Since $\tilde{g}_{D, n}$ occasionally can have negative values, we propose the MAD-type estimator

$$g_{D, n} = \max\{\tilde{g}_{D, n}, 0\} \quad (5)$$

instead of $\tilde{g}_{D, n}$. We use the smoothed version L_{ε} of the pinball loss function $L_{0.5-\text{pin}}$ because we will need in the proof of Theorem 2.2 that the loss function has a Lipschitz continuous derivative, see (18) and (19). This is the price we pay for the unavoidable fact that our estimation cannot be based on the true residuals but on the estimated ones because the distribution P of (X_i, Y_i) is assumed to be unknown in statistical machine learning. Some easy calculations show that the smoothed pinball loss function L_{ε} is convex, Lipschitz continuous with $|L_{\varepsilon}|_1 = 0.5$, has a Lipschitz continuous derivative, and fulfills $0 \leq L_{0.5-\text{pin}}(y, t) - L_{\varepsilon}(y, t) \leq \log(2)\varepsilon < \varepsilon$ for every $(y, t) \in \mathcal{Y} \times \mathbb{R}$ and the risks fulfill, for all $P \in \mathcal{M}_1(\mathcal{Y} \times \mathbb{R})$,

$$0 \leq \mathbb{E}_P L_{0.5-\text{pin}}(Y, f(X)) - \mathbb{E}_P L_{\varepsilon}(Y, f(X)) \leq \mathbb{E}_P |L_{0.5-\text{pin}}(Y, f(X)) - L_{\varepsilon}(Y, f(X))| < \varepsilon.$$

The ε -smoothed version of the pinball loss is actually a re-parametrized logistic loss function $L_{\varepsilon}(y, t) = \varepsilon L_{\text{logistic}}(y/\varepsilon, t/\varepsilon)/2$, see [23, p. 44]. Hence the SVM based on L_{ε} can be calculated by any software which supports the logistic loss. For the illustration purposes in the introduction, we used $\varepsilon = 0.1$ and calculated (3) by Newton-Raphson.

For any loss function L and every measurable $f, g : \mathcal{X} \rightarrow \mathbb{R}$, define the risk

$$\mathcal{R}_{L, P}(f, g) := \mathbb{E}_P L(|Y - f(X)|, g(X)). \quad (6)$$

If the median function $f_{0.5, P}^*$ and the MAD function $g_P^*(x) = \text{MAD}(Y|X = x)$ uniquely exist, then the MAD function g_P^* minimizes $g \mapsto \mathcal{R}_{L_{0.5-\text{pin}}, P}(f_{0.5, P}^*, g)$ over all measurable functions $g : \mathcal{X} \rightarrow \mathbb{R}$, i.e.

$$\mathcal{R}_{L_{0.5-\text{pin}}, P}(f_{0.5, P}^*, g_P^*) = \inf_{g \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L_{0.5-\text{pin}}, P}(f_{0.5, P}^*, g).$$

The following theorem says that $g_{D, n}$ is risk ε -consistent for the MAD function.

Theorem 2.2 *In addition to Assumption 2.1, assume that $\mathbb{E}_P|Y| < \infty$ and that the median function $f_{0.5,P}^* : \mathcal{X} \rightarrow \mathbb{R}$ is almost surely unique. Let L_0 denote the 0.5-pinball loss function and let $\varepsilon > 0$ be the predefined real number in the loss function L_ε . Then, for $n \rightarrow \infty$,*

$$\inf_{g \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L_0,P}(f_{0.5,P}^*, g) + \varepsilon \geq \mathcal{R}_{L_0,P}(f_{0.5,P}^*, g_{D,n}) + o_P(1) = \mathcal{R}_{L_0,P}(f_{L_0,D,\lambda_{1,n}}, g_{D,n}) + o_P(1)$$

if $\lim_{n \rightarrow \infty} \lambda_{1,n} = 0$, $\lim_{n \rightarrow \infty} \lambda_{2,n} = 0$, and $\lim_{n \rightarrow \infty} \lambda_{1,n}^2 \lambda_{2,n}^2 n = \infty$.

Remarks: (i) We assume that the true median function uniquely exists but we do *not* assume that the true MAD function g_P^* uniquely exists. (ii) The value $\mathcal{R}_{L_0,P}(f_{0.5,P}^*, g_{D,n})$ quantifies the expected distance of the estimate $g_{D,n}$ to the absolute values of the *true* residuals $|Y - f_{0.5,P}^*(X)|$; the value $\mathcal{R}_{L_0,P}(f_{L_0,D,\lambda_{1,n}}, g_{D,n})$ quantifies the expected distance of the estimate $g_{D,n}$ to the absolute values of the *estimated* residuals $|Y - f_{L_0,D,\lambda_{1,n}}(X)|$. According to Theorem 2.2, both values asymptotically achieve the infimal risk up to the predefined $\varepsilon > 0$. (iii) The assumption $\lim_{n \rightarrow \infty} \lambda_{1,n}^2 \lambda_{2,n}^2 n = \infty$ is stronger than the standard assumption $\lim_{n \rightarrow \infty} \lambda_{j,n}^2 = \infty$; see [23, Thm. 9.6]. This is plausible because estimating the MAD is burdened with the estimation of a nuisance function (i.e. the unknown median function).

2.2 IQR-type estimation

Let us now consider a linear combination of m SVMs under the Assumption 2.1. As the results follow by straightforward calculations using standard results on SVMs, the proofs are left out.

Let m be a positive integer, $J = \{1, \dots, m\}$, $c = (c_1, \dots, c_m) \in \mathbb{R}^m \setminus \{0\}$, and $(\xi_{j,n})_{n \in \mathbb{N}_0}$ be a sequence of measurable functions into some complete separable metric space E_j equipped with its Borel σ -algebra, $j \in J$. Obviously, $\sum_{j \in J} c_j \xi_{j,n}$ exists and is unique if all $\xi_{j,n}$ exist and are unique. Furthermore, $\sum_{j \in J} c_j \xi_{j,n}$ converges to $\sum_{j \in J} c_j \xi_{j,0}$ in probability (or almost surely or in the L_p sense) if all $\xi_{j,n}$ converge in probability (or almost surely or in the L_p sense) to $\xi_{j,0}$ for $n \rightarrow \infty$.

Now, let $0 < \tau_1 < \dots < \tau_m < 1$. If we either specialize that $\xi_{j,n}$ denotes the support vector machine $f_{L_{\tau_j-\text{pin}},D,\lambda_{j,n}}$ and choose as E_j the RKHS H_j or that $\xi_{j,n}$ denotes the corresponding risk $\mathbb{E}_P L_{\tau_j-\text{pin}}(Y, f_{L_{\tau_j-\text{pin}},D,\lambda_{j,n}}(X))$ and choose $E_j = \mathcal{Y}$, then existence, uniqueness and consistency results for the linear combination of the SVMs or of their risks follow immediately from results valid for each individual SVM, see, e.g., [23, 22, 24] and our Theorem 3.2. Denote the subdifferential (see e.g. [7, Section 5.3]) of the pinball loss function $L_{\tau_j-\text{pin}}$ (with respect to the second argument) by $\partial L_{\tau_j-\text{pin}}$. We then obtain immediately a representer theorem for the linear combination of SVMs because it is well-known that each individual SVM has a representer theorem, i.e. it holds

$$\sum_{j \in J} c_j f_{L_{\tau_j-\text{pin}},P,\lambda_{j,n}} = \sum_{j \in J} -\frac{1}{2\lambda_j} c_j \mathbb{E}_P h_{j,P}(X, Y) \Phi_j(X), \quad (7)$$

where the functions $h_{j,P}$ fulfill

$$h_{j,P}(x, y) \in \partial L_{\tau_j-\text{pin}}(y, f_{L_{\tau_j-\text{pin}},P,\lambda_{j,n}}(x)) \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad (8)$$

see e.g. [23, Thm.5.8, Cor.5.11]. In the same manner we obtain by straightforward calculations bounds for the maximal bias of SVMs and Bouligand influence functions for linear combinations of SVMs, see [4]. Note that SVMs exist even for heavy-tailed distributions which violate the classical assumption $\mathbb{E}_P|Y| < \infty$, which can be shown by using a trick already used by [9] where instead of the original loss function a shifted loss function in the sense $L_{\tau-\text{pin}}^*(y, t) := L_{\tau-\text{pin}}(y, t) - L_{\tau-\text{pin}}(y, 0)$, $y, t \in \mathbb{R}$, is used, see [5]. This only changes the objective function to be minimized, but not the SVM itself.

Example 2.3 *[Estimation of scale functions.] Let $m = 2$, $c = (-1, +1)$, $\tau_1 \in (0, \frac{1}{2})$ and $\tau_2 \in (\frac{1}{2}, 1)$, e.g. $(\tau_1, \tau_2) = (\frac{1}{4}, \frac{3}{4})$ or $(\tau_1, \tau_2) = (0.05, 0.95)$. Then we obtain immediately existence, uniqueness and consistency results for the difference of the two SVMs based on pinball loss functions $L_{\tau_2-\text{pin}}$ and*

$L_{\tau_1-\text{pin}}$, respectively. In other words, if we denote a τ_j -quantile of the conditional distribution of Y given $X = x$ by $f_{\tau_j, \mathbf{P}}^*$, then the following difference of two SVMs

$$f_{L_{\tau_2-\text{pin}}, \mathbf{D}, \lambda_{2, n}} - f_{L_{\tau_1-\text{pin}}, \mathbf{D}, \lambda_{1, n}}$$

yields an estimator for the difference of $f_{\tau_2, \mathbf{P}}^* - f_{\tau_1, \mathbf{P}}^*$. \blacktriangleleft

Example 2.4 [Estimation of asymmetry functions.] Let $m = 3$, $c = (+1, -2, +1)$, $\tau_1 \in (0, \frac{1}{2})$, $\tau_2 = \frac{1}{2}$, and $\tau_3 \in (\frac{1}{2}, 1)$, e.g. $(\tau_1, \tau_2, \tau_3) = (\frac{1}{4}, \frac{1}{2}, \frac{3}{4})$ or $(\tau_1, \tau_2, \tau_3) = (0.05, 0.5, 0.95)$. Then we obtain immediately existence, uniqueness and consistency results for

$$f_{L_{\tau_3-\text{pin}}, \mathbf{D}, \lambda_{3, n}} - 2f_{L_{\tau_2-\text{pin}}, \mathbf{D}, \lambda_{2, n}} + f_{L_{\tau_1-\text{pin}}, \mathbf{D}, \lambda_{1, n}},$$

which gives us an estimator for the difference of

$$(f_{\tau_3, \mathbf{P}}^* - f_{\tau_2, \mathbf{P}}^*) - (f_{\tau_2, \mathbf{P}}^* - f_{\tau_1, \mathbf{P}}^*). \quad (9)$$

Let us now choose $\tau \in (0, \frac{1}{2})$ and $\tau_1 = 1 - \tau_3 = \tau$, e.g. $\tau = 0.05$. Then the function in (9) is zero, if, for all $x \in \mathcal{X}$, the upper conditional quantile $f_{1-\tau, \mathbf{P}}^*(x)$ differs from the conditional median $f_{0.5, \mathbf{P}}^*(x)$ by the same amount than the conditional median $f_{0.5, \mathbf{P}}^*(x)$ differs from the lower conditional quantile $f_{\tau, \mathbf{P}}^*(x)$. Hence the function in (9) or its supremum norm can be used as a quantity to measure the amount of asymmetry of the conditional distribution of Y given $X = x$. \blacktriangleleft

It is well-known that the so-called *crossing problem* can occur in quantile regression and that this problem is *not* specific to SVMs, see [14, p. 55-59]. The crossing problem occurs if, for two quantile levels $\tau_1 < \tau_2$, the *estimated* quantile functions $\hat{q}_{\tau_1}, \hat{q}_{\tau_2}$ are in the wrong order for at least one $x \in \mathcal{X}$, i.e. $\hat{q}_{\tau_1}(x) > \hat{q}_{\tau_2}(x)$. The danger that the crossing problem occurs for a fixed data set is typically small if τ_1 is close to 0 and if τ_2 is close to 1. A numerical method to prevent the crossing problem in kernel based quantile regression was proposed by [25].

2.3 Comparison of MAD-type and IQR-type estimation

From our point of view, it will often depend on the application whether an MAD- or an IQR-type SVM is more appropriate.

We see three advantages of MAD-type estimation. (i) One can estimate the heteroscedasticity of $P(\cdot|x)$ by estimating the conditional median of the absolute residuals $|Y - \hat{f}(x)|$ without estimating *two* conditional quantile functions. Because in most applications the conditional median (or the conditional mean) are estimated anyway, one only needs to compute *one* additional SVM instead of two additional SVMs by the IQR-type approach. (ii) It can happen that the upper and the lower quantile functions are hard to approximate, e.g., they are not in the RKHSs H_1 and H_2 which can easily happen even with the classical Gaussian RBF kernel whose RKHS contains only continuous functions, see [23, Lem. 4.28, Cor. 4.36] whereas the true quantile functions may have jumps. (iii) It can happen that the *difference* of two quantile functions is easy to estimate, e.g. it is constant, linear, or a polynomial of low order, although the quantile functions $f_{\tau_1, \mathbf{P}}^*$ and $f_{\tau_2, \mathbf{P}}^*$ are complicated.

On the other hand, we see three advantages of IQR-type estimation: (i) Greater flexibility by the choice of (τ_1, τ_2) whereas the MAD-type approach is based on estimating *one* conditional quantile (which is here $\tau = \frac{1}{2}$) of the distribution of the *absolute residuals*. (ii) Greater flexibility by choosing different types of kernels or kernels with different kernel parameters for estimating the upper and the lower quantile functions. (iii) The IQR-type approach allows the direct estimation of asymmetry or other quantities of interest for the distribution of Y given $X = x$.

3 Proofs

3.1 L_1 consistency of quantile function estimation by SVMs

The following lemma strengthens [23, Cor. 3.62] in case of the pinball loss function as convergence in probability is replaced by the stronger L_1 -convergence.

Lemma 3.1 *Let L be the pinball loss with $\tau \in (0, 1)$ and let $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be the distribution of (X, Y) . Assume that $\mathbb{E}_P|Y| < \infty$ and that the conditional quantile function $f_{\tau, P}^* : \mathcal{X} \rightarrow \mathbb{R}$ is $P_{\mathcal{X}}$ -a.s. unique. Then, for every $f_n \in L_1(P_{\mathcal{X}})$, $n \in \mathbb{N}$, we have*

$$\lim_{n \rightarrow \infty} \mathcal{R}_{L, P}(f_n) = \inf_{f \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L, P}(f) \quad \Rightarrow \quad \lim_{n \rightarrow \infty} \|f_n - f_{\tau, P}^*\|_{L_1(P_{\mathcal{X}})} = 0.$$

Proof: Define $h_n : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ by $h_n(x, y) = L(y, f_n(x))$ and $h_0 : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ by $h_0(x, y) = L(y, f_{\tau, P}^*(x))$. Define $c := \min\{1 - \tau, \tau\}$ and note that $L(y, t) \geq c|y - t|$ for every $(y, t) \in \mathcal{Y} \times \mathbb{R}$. According to [23, Cor. 3.62], it is already known that $f_n \rightarrow f_{\tau, P}^*$ in probability (w.r.t. $P_{\mathcal{X}}$). Therefore, it follows from the continuity of L that $h_n \rightarrow h_0$ in probability (w.r.t. P). Since

$$\lim_{n \rightarrow \infty} \int |h_n| dP = \lim_{n \rightarrow \infty} \mathcal{R}_{L, P}(f_n) = \inf_{f \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L, P}(f) = \mathcal{R}_{L, P}(f_{\tau, P}^*) = \int |h_0| dP,$$

the sequence $(h_n)_{n \in \mathbb{N}}$, is uniformly integrable; see e.g. [1, Thm. 21.7]. Since

$$|f_n(x)| \leq |y - f_n(x)| + |y| \leq c^{-1}L(y, f_n(x)) + |y| = c^{-1}h_n(x, y) + |y| \quad \forall (x, y, n) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{N},$$

it follows that the sequence f_n , $n \in \mathbb{N}$, is uniformly integrable, too. Hence convergence in probability of f_n , $n \in \mathbb{N}$, implies L_1 -convergence; see e.g. [1, Thm. 21.7]. \blacksquare

The following theorem strengthens [23, Thm. 9.7(i)] as convergence in probability is replaced by the stronger L_1 -convergence. The proof coincides with that of [23, Thm. 9.7(i)] apart from applying Lemma 3.1 instead of [23, Cor. 3.62] and therefore is omitted.

Theorem 3.2 *Let \mathcal{X} be a complete measurable space, $\mathcal{Y} \subset \mathbb{R}$ be closed, L be the pinball loss with $\tau \in (0, 1)$, H be a separable RKHS of a bounded kernel k on \mathcal{X} such that H is dense in $L_1(\mu)$ for all $\mu \in \mathcal{M}_1(\mathcal{X})$, and $\lambda_n \in (0, \infty)$, $n \in \mathbb{N}$, such that $\lim_{n \rightarrow \infty} \lambda_n = 0$ and $\lim_{n \rightarrow \infty} \lambda_n^2 n = \infty$. Let $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be the distribution of (X, Y) and assume that $\mathbb{E}_P|Y| < \infty$ and that the conditional quantile function $f_{\tau, P}^* : \mathcal{X} \rightarrow \mathbb{R}$ is $P_{\mathcal{X}}$ -a.s. unique. Then,*

$$\|f_{L, D, \lambda_n} - f_{\tau, P}^*\|_{L_1(P_{\mathcal{X}})} \rightarrow 0 \quad \text{in probability,} \quad n \rightarrow \infty.$$

3.2 Proof of Theorem 2.2

In order to increase the readability of the proof, a comprehensive notation is needed. Therefore, we define $L_0 := L_{0.5-\text{pin}}$ and, for probability measures $P_1, P_2 \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, we define

$$\begin{aligned} f_{P_1; n} &:= f_{L_0, P_1, \lambda_{1, n}} = \arg \inf_{f \in H_1} \left(\int L_0(y, f(x)) P_1(d(x, y)) + \lambda_{1, n} \|f\|_{H_1}^2 \right), \\ g_{P_1, P_2; n} &:= \arg \inf_{g \in H_2} \left(\int L_\varepsilon(|y - f_{P_1; n}(x)|, g(x)) P_2(d(x, y)) + \lambda_{2, n} \|g\|_{H_2}^2 \right). \end{aligned}$$

In this definition, P_1 and P_2 can also be equal to the empirical measure $D = \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, Y_i)}$, which corresponds to the random sample $D = ((X_1, Y_1), \dots, (X_n, Y_n))$. That is, the estimate $g_{D, n}$ defined in (5) and (3), is given by $g_{D, n} = \max\{g_{D, D; n}, 0\}$ in this notation. We obtain

$$L'_\varepsilon(y, t) := \frac{\partial}{\partial t} L_\varepsilon(y, t) = \frac{1}{2} - \Lambda\left(\frac{y-t}{\varepsilon}\right) \quad \forall y, t \in \mathbb{R}.$$

Since $|\frac{\partial}{\partial y} L_\varepsilon(y, t)| \leq 0.5$ and $|\frac{\partial}{\partial t} L_\varepsilon(y, t)| \leq 0.5$ for every $y, t \in \mathbb{R}$, the following Lipschitz property is fulfilled

$$|L(y_1, t_1) - L(y_2, t_2)| \leq 0.5|y_1 - y_2| + 0.5|t_1 - t_2| \quad \forall y_1, y_2, t_1, t_2 \in \mathbb{R}. \quad (10)$$

An easy calculation shows that (10) implies

$$|\mathcal{R}_{L_\varepsilon, P}(f_1, g_1) - \mathcal{R}_{L_\varepsilon, P}(f_2, g_2)| \leq 0.5\|f_1 - f_2\|_{L_1(P_{\mathcal{X}})} + 0.5\|g_1 - g_2\|_{L_1(P_{\mathcal{X}})} \quad (11)$$

for all $f_1, f_2, g_1, g_2 \in \mathcal{L}_1(\mathcal{P}_{\mathcal{X}})$. Note that, by construction, $0 \leq L_0 - L_\varepsilon \leq \varepsilon$, which implies

$$\mathcal{R}_{L_\varepsilon, \mathcal{P}}(f, g) \leq \mathcal{R}_{L_0, \mathcal{P}}(f, g) \leq \mathcal{R}_{L_\varepsilon, \mathcal{P}}(f, g) + \varepsilon \quad \forall f, g \in \mathcal{L}_1(\mathcal{P}_{\mathcal{X}}). \quad (12)$$

It is obvious from the definition (6) of the risk $\mathcal{R}_{L_0, \mathcal{P}}(f, g)$ that replacing negative values of the function g by 0 reduces the risk. Hence, the definitions imply $\mathcal{R}_{L_0, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g_{\mathcal{D}, n}) \leq \mathcal{R}_{L_0, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g_{\mathcal{D}, \mathcal{D}; n})$ and it follows from (12) that

$$\begin{aligned} & \mathcal{R}_{L_0, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g_{\mathcal{D}, n}) - \inf_{g \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L_0, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g) \leq \\ & \leq \mathcal{R}_{L_\varepsilon, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g_{\mathcal{D}, \mathcal{D}; n}) - \inf_{g \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L_\varepsilon, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g) + \varepsilon. \end{aligned} \quad (13)$$

Applying the triangular inequality yields

$$\mathcal{R}_{L_\varepsilon, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g_{\mathcal{D}, \mathcal{D}; n}) \leq \left| \mathcal{R}_{L_\varepsilon, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g_{\mathcal{D}, \mathcal{D}; n}) - \mathcal{R}_{L_\varepsilon, \mathcal{P}}(f_{\mathcal{P}; n}, g_{\mathcal{P}; n}) \right| + \mathcal{R}_{L_\varepsilon, \mathcal{P}}(f_{\mathcal{P}; n}, g_{\mathcal{P}; n}). \quad (14)$$

Next, define

$$\begin{aligned} \Delta_1^{(n)} &:= \|g_{\mathcal{D}, \mathcal{D}; n} - g_{\mathcal{P}, \mathcal{D}; n}\|_{L_1(\mathcal{P}_{\mathcal{X}})}, \quad \Delta_2^{(n)} := \|g_{\mathcal{P}, \mathcal{D}; n} - g_{\mathcal{P}, \mathcal{P}; n}\|_{L_1(\mathcal{P}_{\mathcal{X}})}, \\ \Delta_3^{(n)} &:= \|f_{0.5, \mathcal{P}}^* - f_{\mathcal{P}; n}\|_{L_1(\mathcal{P}_{\mathcal{X}})}, \quad \Delta_4^{(n)} := \left(\mathcal{R}_{L_\varepsilon, \mathcal{P}}(f_{\mathcal{P}; n}, g_{\mathcal{P}; n}) - \inf_{g \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L_\varepsilon, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g) \right). \end{aligned}$$

Then, it follows from (13), (14), (11), and another application of the triangular inequality that

$$0 \leq \mathcal{R}_{L_0, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g_{\mathcal{D}, n}) - \inf_{g \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L_0, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g) \leq 0.5 \left(\Delta_1^{(n)} + \Delta_2^{(n)} + \Delta_3^{(n)} \right) + \Delta_4^{(n)} + \varepsilon. \quad (15)$$

Each of the four summands $\Delta_1^{(n)}, \dots, \Delta_4^{(n)}$ will be considered separately in the following four parts. In order to prove the theorem, it is enough to show that $\Delta_1^{(n)}$ and $\Delta_2^{(n)}$ converge to 0 in probability (Part 1 and Part 2), that $\Delta_3^{(n)}$ converges to 0 (Part 3), and that the limit superior of $\Delta_4^{(n)}$ is not larger than 0 (Part 4); the terms $\Delta_3^{(n)}$ and $\Delta_4^{(n)}$ are non-stochastic. Note that (11) and Theorem 3.2 imply, for $n \rightarrow \infty$, the convergence in probability of

$$\left| \mathcal{R}_{L_0, \mathcal{P}}(f_{0.5, \mathcal{P}}^*, g_{\mathcal{D}, n}) - \mathcal{R}_{L_0, \mathcal{P}}(f_{L_0, \mathcal{D}, \lambda_{1, n}}, g_{\mathcal{D}, n}) \right| \leq 0.5 \|f_{0.5, \mathcal{P}}^* - f_{L_0, \mathcal{D}, \lambda_{1, n}}\|_{L_1(\mathcal{P}_{\mathcal{X}})} \rightarrow 0.$$

Part 1: For $D = ((X_1, Y_1), \dots, (X_n, Y_n))$, define

$$Q_D = \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, |Y_i - f_{\mathcal{P}; n}(X_i)|)} \quad \text{and} \quad \tilde{Q}_D = \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, |Y_i - f_{\mathcal{D}; n}(X_i)|)}.$$

For every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, define $h_{\mathcal{D}, n}(x, y) = L'_\varepsilon(y, g_{\mathcal{P}, \mathcal{D}; n}(x))$. Then, it follows from the representer theorem [23, Cor. 5.10] that

$$\begin{aligned} & \|g_{\mathcal{D}, \mathcal{D}; n} - g_{\mathcal{P}, \mathcal{D}; n}\|_{H_2} \leq \lambda_{2, n}^{-1} \|\mathbb{E}_{\tilde{Q}_D} h_{\mathcal{D}, n} \Phi_2 - \mathbb{E}_{Q_D} h_{\mathcal{D}, n} \Phi_2\|_{H_2} \leq \\ & \leq \frac{1}{\lambda_{2, n} n} \sum_{i=1}^n \left| h_{\mathcal{D}, n}(X_i, |Y_i - f_{\mathcal{D}; n}(X_i)|) - h_{\mathcal{D}, n}(X_i, |Y_i - f_{\mathcal{P}; n}(X_i)|) \right| \cdot \|\Phi_2(X_i)\|_{H_2}. \end{aligned} \quad (16)$$

According to the boundedness of k_1 and k_2 , we will use the well-known inequalities

$$\|\Phi_j(x)\|_{H_j} \leq \|k_j\|_\infty \quad \forall x \in \mathcal{X} \quad \text{and} \quad \|f\|_\infty \leq \|k_j\|_\infty \|f\|_{H_j} \quad \forall f \in H_j \quad (17)$$

for every $j \in \{1, 2\}$; see [23, p. 124]. Then, the definition of $h_{D,n}$ and the easy to prove Lipschitz property $|L'_\varepsilon(y_1, t) - L'_\varepsilon(y_2, t)| \leq \varepsilon^{-1}|y_1 - y_2|$ for all $y_1, y_2, t \in \mathbb{R}$ imply

$$\begin{aligned} \|g_{D,D;n} - g_{P,D;n}\|_{H_2} &\stackrel{(16,17)}{\leq} \frac{\|k_2\|_\infty}{\lambda_{2,n} n} \sum_{i=1}^n |h_{D,n}(X_i, |Y_i - f_{D;n}(X_i)|) - h_{D,n}(X_i, |Y_i - f_{P;n}(X_i)|)| \\ &\leq \|k_2\|_\infty \lambda_{2,n}^{-1} \sup_t \sup_{x,y} |L'_\varepsilon(|y - f_{D;n}(x)|, t) - L'_\varepsilon(|y - f_{P;n}(x)|, t)| \end{aligned} \quad (18)$$

$$\leq \|k_2\|_\infty \varepsilon^{-1} \lambda_{2,n}^{-1} \sup_{x,y} \left| |y - f_{D;n}(x)| - |y - f_{P;n}(x)| \right| \leq \|k_2\|_\infty \varepsilon^{-1} \lambda_{2,n}^{-1} \|f_{D;n} - f_{P;n}\|_\infty \quad (19)$$

$$\stackrel{(17)}{\leq} \|k_1\|_\infty \|k_2\|_\infty \varepsilon^{-1} \lambda_{2,n}^{-1} \|f_{D;n} - f_{P;n}\|_{H_1}.$$

Next, it follows from the representer theorem [23, Cor. 5.10] that there is an $h_{P,n} \in \mathcal{L}_\infty(\mathcal{X})$ such that $\|h_{P,n}\|_\infty \leq 0.5$ and

$$\|f_{D;n} - f_{P;n}\|_{H_1} \leq \lambda_{1,n}^{-1} \left\| \frac{1}{n} \sum_{i=1}^n (h_{P,n}(X_i, Y_i) \Phi_1(X_i) - \mathbb{E}_P h_{P,n} \Phi_1) \right\|_{H_1}. \quad (20)$$

Define $B := \sup_{x,y} \|h_{P,n}(x, y) \Phi_1(x)\|_{H_1} \leq 0.5 \|k_1\|_\infty$ and fix any $\eta > 0$. Then it follows from (20) and Hoeffding's inequality [28, Chapter 3] that, for $n \rightarrow \infty$,

$$P^n \left(\lambda_{2,n}^{-1} \|f_{D;n} - f_{P;n}\|_{H_1} \geq \eta \right) \leq \exp \left(- \frac{3}{8} \cdot \frac{\eta^2 \lambda_{1,n}^2 \lambda_{2,n}^2 n}{\eta \lambda_{1,n} \lambda_{2,n} B + 3B^2} \right) \rightarrow 0,$$

because $\lim_{n \rightarrow \infty} \lambda_{1,n}^2 \lambda_{2,n}^2 n = 0$. That is, we have shown that $\Delta_1^{(n)} = \|g_{D,D;n} - g_{P,D;n}\|_{H_2}$ converges to 0 in probability w.r.t. P^n .

Part 2: Define $L_{\varepsilon;n}(x, y, t) = L_\varepsilon(|y - f_{P;n}(x)|, t)$. This yields

$$g_{P,P;n} := \arg \inf_{g \in H_2} \left(\int L_{\varepsilon;n}(x, y, g(x)) P_1(d(x, y)) + \lambda_{2,n} \|g\|_{H_2}^2 \right) \quad \forall P_1 \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}).$$

Hence, for $h_{P,n}(x, y) = L'_\varepsilon(|y - f_{P;n}(x)|, t)$, the representer theorem [23, Cor. 5.10] implies that

$$\|g_{P,D;n} - g_{P,P;n}\|_{H_2} \leq \lambda_{2,n}^{-1} \left\| \frac{1}{n} \sum_{i=1}^n h_{P,n} \Phi_2 - \mathbb{E}_P h_{P,n} \Phi_2 \right\|_{H_2}.$$

For $B := \sup_{x,y} \|h_{P,n}(x, y) \Phi_2(x)\|_{H_1} \leq 0.5 \|k_2\|_\infty$, it follows from Hoeffding's inequality [28, Chap. 3] and $\lambda_{2,n}^2 n \rightarrow \infty$ that $\Delta_2^{(n)} = \|g_{P,D;n} - g_{P,P;n}\|_{H_2}$ converges to 0 in probability.

Part 3: Since $\lim_{n \rightarrow \infty} \mathcal{R}_{L_0, P}(f_{P;n}) = \inf_{f \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L_0, P}(f)$ as shown in [23, p. 338], it follows from Lemma 3.1 that

$$\lim_{n \rightarrow \infty} \Delta_3^{(n)} = \lim_{n \rightarrow \infty} \|f_P^* - f_{P;n}\|_{L_1(P_{\mathcal{X}})} = 0. \quad (21)$$

Part 4: For every $g \in H_2$, define the approximation error function (where we use the notation (6))

$$A_g : L_1(P_{\mathcal{X}}) \times \mathbb{R} \rightarrow \mathbb{R}, \quad (f, \lambda) \mapsto \mathcal{R}_{L_\varepsilon, P}(f, g) + \lambda \|g\|_{H_2}^2 - \inf_{g_0 \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L_\varepsilon, P}(f_{0.5, P}^*, g_0).$$

Note that the assumption $\mathbb{E}_P|Y| < \infty$ implies that $|A_g(f, \lambda)| < \infty$ such that A_g is well defined. It follows from the Lipschitz property (10) of L_ε that A_g is continuous for every $g \in H_2$ and, therefore, the map $(f, \lambda) \mapsto \inf_{g \in H_2} A_g(f, \lambda)$ is upper semicontinuous. Hence, (21) implies

$$\limsup_{n \rightarrow \infty} \Delta_4^{(n)} \leq \limsup_{n \rightarrow \infty} \inf_{g \in H_2} A_g(f_{P;n}, \lambda_{2,n}) \leq \inf_{g \in H_2} A_g(f_{0.5, P}^*, 0) = 0,$$

where the last equality follows, because the assumption that H_2 is dense in $L_1(P_{\mathcal{X}})$ guarantees $\inf_{g_0 \in \mathcal{L}_0(\mathcal{X})} \mathcal{R}_{L_\varepsilon, P}(f_{0.5, P}^*, g_0) = \inf_{g \in H_2} \mathcal{R}_{L_\varepsilon, P}(f_{0.5, P}^*, g)$ according to [23, Theorem 5.31]. \blacksquare

References

- [1] H. Bauer. *Measure and integration theory*. Walter de Gruyter & Co., Berlin, 2001.
- [2] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer, Boston, 2004.
- [3] T. T. Cai and L. Wang. Adaptive variance function estimation in heteroscedastic nonparametric regression. *The Annals of Statistics*, 36:2025–2054, 2008.
- [4] A. Christmann and A. Van Messem. Bouligand derivatives and robustness of support vector machines for regression. *J. Mach. Learn. Res.*, 9:915–936, 2008.
- [5] A. Christmann, A. Van Messem, and I. Steinwart. On consistency and robustness properties of support vector machines for heavy-tailed distributions. *Statistics and Its Interface*, 2:311–327, 2009.
- [6] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, 2000.
- [7] Z. Denkowski, S. Migórski, and N. Papageorgiou. *An introduction to nonlinear analysis: Theory*. Kluwer Academic Publishers, Boston, 2003.
- [8] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36:1171–1220, 2008.
- [9] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. 5th Berkeley Symp.*, 1:221–233, 1967.
- [10] P. J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.
- [11] B. Jørgensen. Exponential dispersion models. *J. R. Statist. Soc. B*, 49:127–162, 1987.
- [12] B. Jørgensen. *Theory of Dispersion Models*. Chapman and Hall, London, 1997.
- [13] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11:1–20, 2004.
- [14] R. W. Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.
- [15] R. W. Koenker and G. W. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.
- [16] S. Rüping. *mySVM-Manual*. Department of Computer Science, University of Dortmund, 2000. www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM.
- [17] D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, Cambridge, 2003.
- [18] B. Schölkopf and A. J. Smola. *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.
- [19] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10:1299–1319, 1998.
- [20] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Comput.*, 12:1207–1245, 2000.
- [21] G. Smyth. Generalized linear models with varying dispersion. *J. R. Statist. Soc. B*, 51:47–60, 1989.
- [22] I. Steinwart and A. Christmann. How SVMs can estimate quantiles and the median. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 305–312. MIT Press, Cambridge, MA, 2008.
- [23] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.
- [24] I. Steinwart and A. Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.
- [25] I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola. Nonparametric quantile estimation. *J. Mach. Learn. Res.*, 7:1231–1264, 2006.
- [26] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.
- [27] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Schölkopf, C. J. C. Burges, and A. Smola, editors, *Advances in Kernel Methods–Support Vector Learning*, pages 69–88. MIT Press, Cambridge, MA, 1999.
- [28] V. Yurinsky. *Sums and Gaussian Vectors*, volume 1617 of *Lecture Notes in Mathematics, 1617*. Springer, Berlin, 1995.